DOCUMENT RESUME

ED 440 852 SE 063 471

AUTHOR	Webb, Norman L.
TITLE	Alignment of Science and Mathematics Standards and Assessments in Four States. Research Monograph No. 18.
INSTITUTION	National Inst. for Science Education, Madison, WI.; Council of Chief State School Officers, Washington, DC.
SPONS AGENCY	National Science Foundation, Arlington, VA.
PUB DATE	1999-08-00
NOTE	44p.
CONTRACT	REC-9803080; RED-9452971
AVAILABLE FROM	Wisconsin Center for Education Research, 1025 W. Johnson Street, Madison, WI 53706 ($7.75). Tel: 608-265-9698.
PUB TYPE	Reports - Research (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Academic Standards; *Educational Assessment; Educational Change; Elementary Secondary Education; Evaluation Methods; Instructional Materials; *Item Analysis; *Mathematics Education; *Science Education

ABSTRACT

Reviewers analyzed the alignment of assessments and standards in mathematics and science from four states at a four-day institute conducted June 29 through July 2, 1998. Six reviewers compared the match between assessment items and standards in mathematics and seven compared the match in science. Data from these analyses were processed and used to judge the degree of alignment on four criteria: categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance of representation. The analyses indicated that the standards of the four states varied in what content students were expected to know, the level of specificity at which expectations were expressed, and organization. Nearly all of the sixteen assessment instruments reviewed incorporated some constructed-response items. Only one mathematics assessment for grade 10 from one state consisted solely of multiple-choice items. The items in three science and two mathematics assessments analyzed from one state were evenly divided between multiple-choice and constructed-response items. Assessments from the other three states included from 80% to 90% multiple-choice items. Alignment between assessments and standards varied across grade levels, content areas, and states without any discernable pattern. Assessments and standards of three of the four states satisfied the categorical concurrence criterion. This criterion, the most common conception of alignment, required the assessment and standards to include the same content topics. Alignment was found to be the weakest on the depth-of-knowledge consistency and range-of-knowledge correspondence criteria. Generally, assessment items required a lower level of knowledge and did not span the full spectrum of knowledge as expressed in the standards. However, for the knowledge and skills identified in the standards and addressed by the assessments, generally the assessment items were evenly distributed. A major goal of this study was to develop a valid and reliable process for analyzing the alignment among standards and assessments. The process did produce credible results that distinguished among the different attributes of alignment and detected specific ways that alignment could be improved. Issues that did arise from an analysis of the process indicated that reviewers could benefit from more training at the beginning of the institute. Reviewers also needed more

Reproductions supplied by EDRS are the best that can be made
from the original document.

clarification of the four depth-of-knowledge levels and more explicit rules for assigning an assessment item to more than one statement of expectation. (Contains 15 references.) (CCM)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 440 852

National Institute for Science Education

University of Wisconsin-Madison • National Center for Improving Science Education

Council of Chief State School Officers
Washington, DC

Research Monograph No. 18

Alignment of Science and Mathematics Standards and Assessments in Four States

Norman L. Webb

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

Funded by the
National Science Foundation

National Institute for Science Education (NISE) Publications

The NISE issues papers to facilitate the exchange of ideas among the research and development community in science, mathematics, engineering, and technology (SMET) education and leading reformers of SMET education as found in schools, universities, and professional organizations across the country. The NISE Occasional Papers provide comment and analysis on current issues in SMET education including SMET innovations and practices. The papers in the NISE Research Monograph series report findings of original research. The NISE Conference and Workshop Reports result from conferences, forums, and workshops sponsored by the NISE. In addition to these three publication series, the NISE publishes Briefs on a variety of SMET issues.

The alignment study was supported by a grant to the Council of Chief State School Officers from the National Science Foundation (Award Number REC-9803080) and by the National Institute for Science Education under a cooperative agreement between the National Science Foundation and the UW–Madison (Cooperative Agreement No. RED-9452971). At UW–Madison, the National Institute for Science Education is housed in the Wisconsin Center for Education Research and is a collaborative effort of the College of Agricultural and Life Sciences, the School of Education, the College of Engineering, and the College of Letters and Science. The collaborative effort is also joined by the National Center for Improving Science Education, Washington, DC. Any opinions, findings, or conclusions are those of the author and do not necessarily reflect the view of the supporting agencies.

Research Monograph No. 18

ALIGNMENT OF SCIENCE AND MATHEMATICS STANDARDS AND ASSESSMENTS IN FOUR STATES

Norman L. Webb

National Institute for Science Education
University of Wisconsin-Madison

Council of Chief State School Officers
Washington, DC

August, 1999

BEST COPY AVAILABLE

About the Author

Norman L. Webb, senior research scientist with the Wisconsin Center for Education Research, is a mathematics educator and evaluator who is co-team leader of the Institute's Systemic Reform Team, rethinking how we evaluate mathematics and science education, while focusing on the National Science Foundation's Systemic Initiatives reform movement. His own research has focused on assessment of students' knowledge of mathematics and science. Webb also directs evaluations of curriculum and professional development projects.

Acknowledgements

The author acknowledges the assistance of John Smithson, researcher and data analyst; Margaret Powell, editor, and Lynn Lunde, secretary, for their assistance in the preparation of this monograph. In addition, the Alignment Institute participants were helpful in the preparation of this monograph: David Bahna, Science Education Assessment, South Carolina Department of Education; Rolf Blank, Director of Education Indicators, Council of Chief State School Officers; Jennifer Falls, Mathematics Education, Louisiana Department of Education; Mary Gromko, Science Education, Colorado Department of Education; Michael Kestner, Mathematics Education, North Carolina Department of Education; Gerald Kulm, Mathematics Education, American Association for the Advancement of Science; Michael Lower, Mathematics Education Assessment, South Carolina Department of Education; Megan Martin, Science Assessment Consultant, California; Curtis McKnight, Department of Mathematics, University of Oklahoma; Andrew C. Porter, Director, Wisconsin Center for Education Research, University of Wisconsin-Madison; Harold Pratt, Science Education, National Research Council, Center for Science, Mathematics, and Engineering Education; Senta Raizen, Director, National Center for Improving Science Education; Eleanor Sanford, Assessment, North Carolina Department of Public Instruction; and Linda Wilson, School of Education, Mathematics, University of Delaware.

BEST COPY AVAILABLE

Table of Contents

Figure and Tables v

Executive Summary vii

Summary Report 1

Introduction 1
- Initial Methodology Developed at the Institute for the Analysis of Alignment Criteria 3

Alignment Criteria Used for This Analysis 6
- Categorical Concurrence 7
- Depth-of-Knowledge Consistency 7
- Range-of-Knowledge Correspondence 8
- Balance of Representation 8

State Reports on Alignment 9

Findings: Alignment of States' Standards and Assessments 11
- Categorical Concurrence 11
- Depth-of-Knowledge Consistency 12
- Range-of-Knowledge Correspondence 16
- Balance of Representation 17
- Reviewer Agreement in Coding 18
- Summary of Findings on Alignment Criteria 19

Findings: The Process for Studying Alignment 19
- Reviewers and Their Training 20
  - Coding Process 21
- Levels for Determining Depth of Knowledge 21
  - Coding Procedures 23
- Limitations of the Alignment Analysis 25
- Setting an Acceptable Level for Alignment Criteria 26

Recommendations for Improving the Process 27

Conclusions 27

References 29

Appendix A: Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education 31

Appendix B: Sample of Tables Included in Each State Report: State A Grade 8 Science 33

Figure and Tables

Figure 1. Example of two assessment items with the same stem, but rated at different depth-of-knowledge levels 24

Table 1. Percent of multiple-choice items of total assessment by state, content area, and grade 10
Table 2. Summary of alignment analysis for four states in science and mathematics 13
Table 3. Average percent across standards of reviewers' agreement on acceptable level for criterion 18

Executive Summary

Reviewers analyzed the alignment of assessments and standards in mathematics and science from four states at a four-day institute conducted June 29 through July 2, 1998. Six reviewers compared the match between assessment items and standards in mathematics and seven compared the match in science. Data from these analyses were processed and used to judge the degree of alignment on four criteria: categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance of representation.

The analyses indicated that the standards of the four states varied in what content students were expected to know, the level of specificity at which expectations were expressed, and organization. Nearly all of the sixteen assessment instruments reviewed incorporated some constructed-response items. Only one mathematics assessment for grade 10 from one state consisted solely of multiple-choice items. The items in three science and two mathematics assessments analyzed from one state were evenly divided between multiple-choice and constructed-response items. Assessments from the other three states included from 80% to 90% multiple-choice items.

Alignment between assessments and standards varied across grade levels, content areas, and states without any discernable pattern. Assessments and standards of three of the four states satisfied the categorical concurrence criterion. This criterion, the most common conception of alignment, required the assessment and standards to include the same content topics. Alignment was found to be the weakest on the depth-of-knowledge consistency and range-of-knowledge correspondence criteria. Generally, assessment items required a lower level of knowledge and did not span the full spectrum of knowledge as expressed in the standards. However, for the knowledge and skills identified in the standards and addressed by the assessments, generally the assessment items were evenly distributed.

A major goal of this study was to develop a valid and reliable process for analyzing the alignment among standards and assessments. The process did produce credible results that distinguished among the different attributes of alignment and detected specific ways that alignment could be improved. Issues that did arise from an analysis of the process indicated that reviewers could benefit from more training at the beginning of the institute. Reviewers also needed more clarification of the four depth-of-knowledge levels and more explicit rules for assigning an assessment item to more than one statement of expectation.

Summary Report

Establishing alignment of standards and assessments alone is not enough for attaining the full impact of standards-based reform, but it is an early indicator that helps assure a state's standards and assessments will reach their full potential. Establishing the degree to which assessments are aligned with standards is not easy. This analysis demonstrates one process for quantifying the alignment between standards and assessments, using specific criteria. It summarizes the findings from an alignment analysis of the standards and assessments from four states conducted with the participation of experts in science and mathematics education as reviewers. Four companion reports, one for each state, describe in more detail the analysis and findings from this study. Along with determining the alignment of state standards and assessments, a second important purpose of the study was to refine the process for analyzing alignment. Each of the four states volunteered to have their standards and assessments analyzed for two or three grade levels in mathematics and in science. Throughout this document and the four companion reports, the states are identified as State A, State B, State C, and State D to protect their identity.

Introduction

Alignment is not a new phenomenon, but has been studied for a number of years. What has changed is the nature of the assessments, expectations, and other system components to be aligned and the stakes for achieving alignment. In the 1960s, analyses were performed on assessment tasks and behavioral objectives as part of the mastery-learning movement (Cohen, 1987; Carroll, 1963). Exact alignment was achieved if the assessment tasks were equivalent to the instructional tasks. Learning goals were partitioned into narrowly defined behavioral objectives. Domains of all possible test items were specified for each behavioral objective. Content analysis by expert panels remains the primary technique for judging alignment between standards and assessments. But with the advent of standards-based education, systemic reform (Smith & O'Day, 1991), and criterion-referenced tests, judging alignment has become more complex and requires more systematic procedures. The underlying assumptions regarding the assessments, such as norm-referenced tests and normally distributed achievement, can result in misalignment with standards that are targeted for all students (Baker, Freeman, & Clayton, 1991). Educators increasingly recognize that if system components are not aligned, the system will be fragmented, will send mixed messages, and will be less effective (Consortium for Policy Research in Education, 1991; Newmann, 1993; Spillane, 1998). But in addition to conceptual reasons for assuring alignment, states are also faced with legal reasons. The Improving America's Schools Act explained how assessments are to relate to standards: ". . . such assessments (high quality, yearly student assessments) shall . . . be aligned with the State's challenging content and student performance standards and provide coherent information about student attainment of such standards . . ." (U.S. Congress, 1994, p. 8). The U.S. Department of Education's explanation of the Goals 2000: Educate America Act and the Elementary and Secondary Education Act, which includes Title I, indicated alignment of curriculum, instruction, professional development, and assessments as a key performance indicator for states, districts, and schools striving to meet challenging standards. Within the changing climate of what we know about what works in education and the increasing mandates and pressures on education systems, alignment has become critical to a full understanding of how systems function. This study was directed toward refining procedures for determining degrees of alignment so that they are more standardized and

useful in order for states and districts to better understand the agreement between standards and assessments

Alignment of standards for student learning and assessments for measuring students' attainment of these standards is an essential attribute for an effective standards-based education system. Alignment is defined as the degree to which standards and assessments are in agreement and serve in conjunction with one another to guide the system toward students learning what they are expected to know and do. As such, alignment is a quality of the relationship between standards and assessments and not an attribute of any one of these two system components. As a relationship between two or more system components, alignment can be determined by using the multiple criteria described in detail in a National Institute for Science Education (NISE) research monograph, *Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education* (Webb, 1997).

A four-day Alignment Analysis Institute was conducted June 29 through July 2, 1998. Sixteen people, including state content specialists, state assessment consultants, content experts, and researchers, attended the institute, which was coordinated by the Council of Chief State School Officers (CCSSO) with the cooperation of the National Institute for Science Education (NISE). Prior to this institute, most participants attended a one-day meeting in Washington, DC, on April 29, to be introduced to the process and to the alignment criteria to be used at the institute. At the summer institute, six of the participants rated mathematics standards and assessments; seven rated science standards and assessments; and three coordinated the process. Four states volunteered to have their mathematics standards and assessments analyzed for alignment for two or three grade levels. Three of these states agreed to have their science standards and assessments analyzed for two or three grade levels.[1]

A major goal of the institute was to develop a systematic process and analytic tools for judging the alignment between standards and assessments based on the criteria developed in conjunction with CCSSO and NISE (Webb, 1997) that are listed in Appendix A. Reviewers were not given lengthy training in applying the criteria, but were expected to help perfect the process over the duration of the institute. One outcome of the institute is a refined process that can be used under more controlled conditions to make a judgment on the alignment of standards and assessments. Reviewers were instructed to attend to the alignment between the state standards and assessments. There was no opportunity for reviewers to offer their opinions on either the quality of the standards or of the assessment activities/items. The results produced from the institute pertain only to how the state standards and the state assessments are in agreement; they do not serve as external verification of the general quality of a state's standards or assessments. The results of the Alignment Analysis Institute do provide expert judgment about alignment, independent of any of the participating states, by those who are very familiar with state and

[1] For state C grades 4 and 8 science only a sample of 14 items for each grade were available for review. Sixteen analyses compared at least some assessment items with the state's standards. Fourteen of these analyses used the complete assessment instrument.

national standards.[2] When reviewers did vary in their judgments, using averages lessened the error that might result from any one reviewer.

This report describes the results of an alignment study of standards and grade level tests in mathematics and science for three states and in mathematics only for one state. The study addressed specific criteria related to the content agreement between the state standards and grade level assessments. Four criteria received major attention: categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance of representation. Other criteria such as articulation across grades and ages, equity and fairness, and pedagogical implications were given less emphasis. Wixcon and her colleagues (Wixcon, Fisk, Dutro, & McDaniel, 1999) have successfully applied the four criteria used in this analysis in the context of reading.

Initial Methodology Developed at the Institute for the Analysis of Alignment Criteria

Prior to analyzing the documents, the reviewers were only given general instructions and broad definitions for the depth-of-knowledge levels required to satisfy a standard and to successfully complete an assessment activity. One purpose for conducting these alignment studies is to better specify what training reviewers need if they are to validly code assessment activities and standards. Reviewers were given the following levels to judge depth of knowledge for both mathematics and science:

Levels

1. Recall

Recall of a fact, information, or procedure.

2. Skill/Concept

Use of information, conceptual knowledge, procedures, two or more steps, etc.

3. Strategic Thinking

Requires reasoning, developing a plan or sequence of steps; has some complexity; more than one possible answer; generally takes less than 10 minutes to do.

4. Extended Thinking

Requires an investigation; time to think and process multiple conditions of the problem or task; and more than 10 minutes to do non-routine manipulations.

Reviewers within a content area were encouraged to refine these levels or to add greater clarification, providing they all came to some agreement. One of the intended outcomes for this alignment study is greater clarity for the levels. The revised levels are given in this report.

---

[2] Averages across reviewers and for each standard were computed for variables representing the relationship between standards and assessments. These averages were compared with predefined levels to determine whether alignment was acceptable for each of four criteria.

Different states use different terminology to label expectations for what students are to know and do. Some states label the large categories of student expectations as "strands." Other states call these expectations "competency goals." Still others refer to state expectations as "benchmarks." To improve the interpretation of results, the same convention was used in this analysis to label the different levels of expectations. The term standards refers to the most general expectations for a grade and content area. The number of standards in the four states that participated in this analysis ranged from four to ten. Goal refers to the next level of specificity of expectations. Generally the set of goals for a standard covers the full range of knowledge specified by the standard. The number of goals for a standard in this analysis went as high as 20. Objective refers to the third level of specificity. Objectives further delineate expectations stated as a goal. The number of expectation levels can vary. In this analysis, a maximum of three levels of expectations was included. If a state only used two levels of expectations, then the most general level is called standards and the second level is called objectives.

Prior to the Alignment Analysis Institute, reviewers were sent copies of the standards and were asked to become familiar with them. At the institute, reviewers as a group began by assigning a depth-of-knowledge level for each objective. Achieving one objective could require students to know the content at more than one depth-of-knowledge level.[3] The assigned level was to represent the highest level of knowledge expected for a student to satisfactorily demonstrate the attainment of the objective. All of the reviewers reached consensus on the assigned level for each objective through deliberation as a group. This activity served two purposes. First, reviewers became more familiar with what students were expected to know and do for each objective. Second, the assigned levels were necessary in order to compare the depth-of-knowledge levels of individual assessment items/activities in the analysis.

Reviewers recorded the depth-of-knowledge level for each objective on a coding matrix prepared prior to the institute. The coding matrix listed all of the objectives for student learning for each standard. These expectations were listed in rows in the same order using the same organization as that used in the state's standards document. For each standard, in sequence, the first row listed the standard, the second row a goal, and the third and subsequent rows objectives. Each standard, goal, and objective was assigned a unique numerical-alpha code.

One column on the coding matrix represented one assessment item/activity. Individual reviewers read each assessment item and assigned it a depth-of-knowledge level. Each reviewer then wrote this depth-of-knowledge level code in the item/activity's column in each row of an objective if a student's response to the item/activity provided information about what the student knew or could do with respect to the objective. Each objective coded for an item was called a hit. Multiple hits were allowed for any one assessment item/activity. Initially, reviewers were not given specifications about limits on the number of hits for any one assessment activity/item. After discussion with other reviewers following the coding of each test, reviewers developed more refined guidelines for multiple hits. This had the effect, as the reviewers gained more experience, of reducing noticeably the number of instances that reviewers marked multiple hits for any one item/activity. The number of multiple hits was one source of variation among

[3] Objective as used in this analysis should not be confused with a behavioral objective designed to express one specific behavior and one depth-of-knowledge level.

reviewers. Reviewers did converge in the number of multiple hits as they became more familiar with the process and developed agreed-upon rules.

Reviewers were asked to code the assessment items/activities independently for each test, with little or no interaction. After all of the reviewers completed coding the instruments, they were asked to select a sample of items and compare their results. The primary purpose of this discussion was to improve the reliability among the reviewers in coding assessment items/activities on the next and subsequent instruments. Reviewers could make changes as they calibrated their work with the other reviewers if they felt it was appropriate. Reviewers discussed both what items/activities were assigned to what objectives and the depth-of-knowledge code assigned to each item.

States included in their assessments both multiple-choice items and constructed-response activities; reviewers did not distinguish between the two formats in coding an assessment item/activity during the coding process. Reviewers assigned each item and activity a depth-of-knowledge level and then recorded in the column for that item the number representing this level in the cell corresponding to each objective most associated with the assessment item/activity. Multiple-choice and constructed-response items both had a range in depth-of-knowledge levels and those of each format could correspond to more than one objective.

The codings for all of the reviewers were entered on a spreadsheet to compute summary statistics. For each assessment instrument and standards document, the codes for each reviewer were tabulated by the frequency of hits and the depth-of-knowledge levels for the hits. Data for all of the objectives for one standard were aggregated or listed as a profile for each standard. The results were reported for each standard.

Statistics for each standard were computed on four alignment criteria for content focus: categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance of representation. The mean number of hits was used to judge the categorical concurrence between the assessment instrument and the standards. The frequencies of hits aggregated across the objectives for each standard and by the depth-of-knowledge levels were used to judge depth-of-knowledge consistency by considering the percentage of hits that were below, at the same level as, or above the level for the objective. The percentage of the objectives hit within a standard was used to judge the range-of-knowledge correspondence within a standard. The distribution of the hits among the objectives for a standard with at least one hit was used to compute the balance of representation for a standard. This analysis is based on the assumption that the set of objectives for a standard spans the entire domain of knowledge and skills a student should demonstrate to fully meet the standard, an assumption not always made. Reviewers were asked for their comments on other alignment criteria that included articulation across grades, pedagogical implications, and equity. Some offered their comments on these criteria, but because of strong time pressures, systematic procedures were not used to gather information on these criteria. Reviewers reacted to the overall process and made suggestions in a debriefing session held at the end of the institute.

All of the statistics were computed for each reviewer. The mean for each statistic was computed using the results for only the reviewers who completed coding all of the items—i.e., at least two

reviewers, and up to seven for some assessments. The mean among reviewers on each statistic is a reasonable approximation for the summary information that lessens the error of any one reviewer in coding. Of course, statistics based on the coding by a greater number of reviewers will be more accurate. Standard deviations, reported along with the mean, provide one indication of the variation among reviewers. Of course, the total number of objectives and the total number of hits for a standard also have to be considered in interpreting the significance of the variation among reviewers.

Alignment Criteria Used for This Analysis

This analysis judged the alignment between the standards and the assessment using four criteria. For each criterion, an acceptable level was defined based on what would be required to assure that students have met the standards. A standard, the most general statement of expectations, was used as the unit of analysis in judging the alignment on each criterion. All of the statistics comparing the agreement between the set of standards and an assessment were computed for each standard. The analysis concluded with a judgment of whether or not there was acceptable alignment on each of the four criteria for each standard. What was considered as an acceptable level was based on specific assumptions. The acceptable levels used in this analysis should be considered as advisory and illustrative, but not absolute. A state may have reasons for setting the acceptable level for criteria higher or lower than specified in this analysis. Factors that can influence what an acceptable level is include the cutoff score for proficient work, the breadth of content coverage in a standard, and time for testing. This report explicitly states what assumptions were made in setting criteria for acceptable alignment.

In evaluating whether an acceptable level was attained on a standard for each of the four criteria, no distinction was made if one assessment item/activity had multiple hits (corresponded to more than one objective). For all four states, the analyses gave equal credit to each hit. Also, nearly all assessment items reviewed were scored as right or wrong. Only a very few items were scored with a rubric that had a possible point value greater than one. There were only a few consequences to these conditions. Allowing one assessment item to correspond to more than one objective improved the likelihood that the assessment and standards would satisfy the requirements for an acceptable level on categorical concurrence and range-of-knowledge correspondence criteria. Because nearly all items were scored as right or wrong, treating all of the assessment items as the same, regardless of their format, had essentially no effect on the results of the analysis.

Some reviewers judged that a few assessment items did not measure any of the content expressed in the state's standards. These items were not included in the analysis for that reviewer. Excluding these "maverick" items, as judged by some reviewers, reduced the total number of hits for the reviewer, which then lowered the mean number of hits across the reviewers. Reducing the number of hits made it more difficult to attain acceptable levels on categorical concurrence and range-of-knowledge correspondence criteria. A statistic could have been computed to represent the percent of items on the assessment that corresponded to the standards. But because nearly all assessment items related to at least some objectives, such a statistic was considered to be less informative and was not computed. Instead, if reviewers found items that did not correspond to any content in the standards, this was noted in the written report.

Reviewers also found assessment items that corresponded to the general goal or standard, but not to any of the objectives. These "generic" items were included in the analysis by adding a generic objective encompassing all material in the goal statement not covered by the objectives. This action increased the total number of objectives for a standard and was noted in the text of the report. The existence of generic items indicated an omission in the standards in that the stated objectives did not fully span all of the content knowledge represented in the goal or standard.

Categorical Concurrence

One aspect of alignment between standards and assessments is if both address the same content categories. The categorical concurrence criterion provides a very general indication if both documents incorporate the same content. ***The criterion of categorical concurrence between standards and assessment is met if the same or consistent categories of content appear in both documents.*** This criterion was judged by determining whether the assessment included items measuring content from each standard.

The analysis assumed that the assessment had to have at least six items measuring content from a standard in order for there to be an acceptable categorical concurrence between the standard and the assessment. The number of items, six, is based on estimating the number of items that could produce a reasonably reliable scale for estimating students' mastery of content on that scale. Of course, many factors have to be considered in determining what a reasonable number is, including the reliability of the scale, the mean score, and cutoff score for determining mastery. Using a procedure developed by Subkoviak (1988) and assuming the cutoff score is the mean and the reliability of one item is .1, it was estimated that six items would produce an agreement coefficient between two equivalent test administrations of at least .63. This indicates that about 63% of the group would be consistently classified as masters or non-masters on the basis of two equivalent test administrations. The agreement coefficient would improve to .77 if the cutoff score is increased to one standard deviation from the mean and to .88, with a cutoff score of 1.5 standard deviations from the mean. None of the four states included in the analysis reported student results by standards or required students to achieve a specified cutoff score on assessment scales related to a standard. If a state did do this, then the state would want a higher agreement coefficient than .63. Six items were assumed as a *minimum* for an assessment scale measuring content knowledge related to a standard, and as a basis for making some decisions about students' knowledge of that standard. A concrete example may help to clarify the rationale. If the mean for six items is 3 and the standard deviation is one, then a cutoff score set at 4 would produce an agreement coefficient of .77. Any fewer items with a mean of one-half of the items would require a cutoff that would only allow a student to miss one item. This would be a very stringent requirement, considering a reasonable standard error of measurement, on the scale.

Depth-of-Knowledge Consistency

Standards and assessments can be aligned not only on the category of content covered by each, but also on the basis of the complexity of knowledge required by each. ***Depth-of-knowledge consistency between standards and assessment indicates alignment if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know***

and do as stated in the standards. The acceptable level for a standard on this criterion is directly related to what is considered passing work on the assessment scale for that standard. For consistency to exist between the assessment and the standard, as judged in this analysis, at least 50% of the items corresponding to an objective had to be at or above the level of knowledge of the objective. Fifty percent, a conservative acceptable level, is based on the assumption that most cutoff points on tests require students to answer correctly more than half of the items to attain a passing score. If at least 50% of the assessment items are required to be at or above the corresponding objectives, then students would have to answer correctly at least one of these items. For example, assume an assessment included six items related to one standard and students were required to answer correctly four of those items to be judged proficient—i.e., 67% of the items. If three, 50%, of the six items were at or above the depth-of-knowledge level of the corresponding objectives, then to achieve a proficient score a student would be required to answer correctly at least one item at or above the depth-of-knowledge level of one objective. Some leeway was used in this analysis on this criterion. If a standard had between 40% to 50% of its corresponding items at or above the depth-of-knowledge levels of the objectives, then it was reported that the criterion was "weakly" met.

Range-of-Knowledge Correspondence

For standards and assessments to be aligned, the breadth of knowledge on both should be comparable. *The range-of-knowledge criterion is used to judge whether a comparable span of knowledge expected of students by a standard is the same as, or corresponds to, the span of knowledge that students need in order to correctly answer the assessment items/activities. The criterion for correspondence between span of knowledge for a standard and the assessment considers the number of objectives within the standard with at least one related assessment item/activity.* At least 50% of the objectives for a standard had to have at least one related assessment item in order for the alignment on this criterion to be judged acceptable. This cutoff for acceptance is based on the assumption that students' knowledge should be tested on content from over half of the domain of knowledge for a standard. This assumes that each objective for a standard has equal weight and that the set of objectives spans the knowledge needed to attain the standard. Depending on the balance in the distribution of items and the need to have a low number of items related to any one objective, the requirement that assessment items need to be related to more than 50% of the objectives for a standard increases the likelihood that students will have to demonstrate knowledge on more than one objective per standard to achieve a minimal passing score. As with the other criteria, a state may choose to make the acceptable cutoff point on this criterion more rigorous by requiring an assessment to include items related to a greater number of the objectives. However, any restriction on the number of items included on the test will place an upper limit on the number of objectives that can be assessed. Range-of-knowledge correspondence is more difficult to attain if the content expectations are partitioned among a greater number of standards and a large number of objectives.

Balance of Representation

In addition to comparable depth and breadth of knowledge, aligned standards and assessments require the knowledge to be distributed equally in both. The range-of-knowledge criterion only considers the number of objectives within a standard hit (a standard with a corresponding item),

but does not take into consideration how the hits (or assessment items/activities) were distributed among these objectives. *The balance-of-representation criterion is used to indicate the extent to which items are evenly distributed across objectives.* An index is used to judge the distribution of assessment items. This index only considers the objectives for a standard that have at least one hit—i.e., one related assessment item/objective. The index is computed by considering the difference in the proportion of objectives and the proportion of hits assigned to the objective. An index value of 1 signifies perfect balance and is obtained if the hits (items/assessment) related to a standard are equally distributed among the objectives for the given standard. If 12 objectives for a standard are hit and there are 24 hits, then perfect balance (a value of 1) would be achieved if each objective had two hits. Index values that approach 0 signify that a large proportion of the hits (items/assessment) were on only one or two of all of the objectives hit. Depending on the number of objectives and the number of hits, a unimodal distribution (most items related to one objective and only one item related to each of the remaining objectives) has an index value of less than .5. A bimodal distribution has an index value of around .55 or .6. Index values of .7 or higher indicate that items/activities are distributed among all of the objectives at least to some degree (e.g., every objective has at least two items) and is used as the acceptable cutoff point on this criterion.

State Reports on Alignment

This analysis produced a report on the alignment of standards and assessments for each of the four states that volunteered to participate in the study (Webb, 1999a, b, c, & d). Each state report describes the state's standards, their organization, and the assessments for those grades in science and mathematics in the alignment study. These reports then describe and discuss the degree to which the standards and assessments are aligned on each of the four criteria by content area and grade level. Each analysis reports in four tables the specific attributes for the standards and assessments and their agreement in one content area for one grade level. The four tables produced for State A grade 8 science are included as Appendix B to give a sample of the information included in each state report.

Even though item format was not taken into consideration in the analysis, format is important in considering alignment. This analysis only used four of several criteria that can be used to study the alignment between standards and assessments (Webb, 1997). Structure-of-knowledge comparability, another content-focus criterion (see Appendix A), attends to the appropriateness of the format of assessment items as compared to how students are to understand the relationship among ideas and the connection among concepts and procedures. A thorough study of alignment between standards and assessments should attend carefully to item format and other ways the structure of knowledge can be represented both in the assessments and standards. Item format did have some relevance to the analysis employed in this study. Non-multiple-choice items generally require more time and can reduce the total number of items on the assessment. A lower number of items on an assessment makes it more difficult for the standards and assessments to attain an acceptable level for the categorical concurrence and range-of-knowledge criteria. For example, State D had from 14 to 24 fewer items in the same grade level and content areas as other states (Table 1). This was due, in part, to State D incorporating a high number of constructed-response items.

Table 1
Percent of Multiple-Choice Items of Total Assessment by State, Content Area, and Grade

Content Area	Grade	Total Number of Assessment Items	Percent of Multiple-Choice Items
		N	%
		State A	
Science	3	44	86
	8	70	86
Mathematics	3	50	76
	6	61	83
		State B	
Mathematics	4	86	93
	8	86	93
	10	70	100
		State C	
Mathematics	4	68	84
	8	74	81
		State D	
Science	3	50	50
	7	49	51
	10	46	54
Mathematics	4	54	59
	8	51	61

The assessments of the four states included both multiple-choice items and constructed-response activities. State B mathematics assessments in grades 4, 8, and 10 had the highest percentage of multiple-choice items of the four states (Table 1). Of a total of 86 items for grade 4, 86 items for grade 8, and 70 items for grade 10, the proportions of multiple-choice items were 93%, 93%, and 100% respectively. State D assessments in mathematics and science had the lowest percentage of multiple and fixed-choice items, ranging from 50% to 61%. In science, the grade 3 assessment had 50 items (50% multiple- or fixed-choice), the grade 7 assessment had 49 items (51% multiple-choice), and the grade 10 assessment had 46 items (54% multiple-choice). In mathematics, State D's grade 4 assessment had 54 items (59% multiple-choice) and the grade 8 assessment had 51 items (61% multiple-choice). The total number of items and the proportion of multiple-choice items for States A and C fell in between. State A science assessments in grades 3 and 8 included 44 and 70 items, respectively. Of these items for both grades, 86% were multiple choice. State A mathematics assessments in grades 3 and 6 included 50 and 61 items, respectively. Of these, grade 3 had 76% and grade 6 had 83% multiple-choice items. State C mathematics assessments in grades 4 and 8 included 68 and 74 items, respectively. Of these, grade 4 had 84% and grade 8 had 81% multiple-choice items.

Findings: Alignment of States' Standards and Assessments

Alignment of standards and assessments varied among the four states included in the analysis and between content areas and grade levels within each state. State B was judged to have the highest degree of alignment of the four states and for the content area and grade level analyzed (Table 2).

Table 2 displays a summary of the findings for the four states. The table reports, for each state, content area, and grade level for which an analysis was performed, the percent of standards with an acceptable level of alignment with the assessment for each of the four criteria. More detailed information is given in the reports for each state as illustrated in the complete set of tables for one analysis given in Appendix B. In the summary information reported in Table 2, the standards and assessment were judged to be fully aligned on the criterion if an acceptable level was attained for all of the standards (100%); highly aligned if an acceptable level was attained for 70% to 99% of the standards; partially aligned if an acceptable level was attained for 50% to 69% of the standards; and poorly aligned if an acceptable level was attained for less than 50% of the standards.

For State B, the analysis was conducted for only mathematics and for three grades. The standards and assessments were fully aligned for all three grades on categorical concurrence, highly aligned for one of the three grades on depth-of-knowledge consistency (grade 8, 71%), acceptably or highly aligned for two grades on range-of-knowledge correspondence (grades 8 and 10, 86% and 100%), and acceptably or highly aligned on balance of representation for all three grade levels (86%, 71%, and 100%) on balance of representation. None of the four standards for grade 10 attained an acceptable degree of alignment for the depth-of-knowledge consistency. For each of the grade 10 standards, less than 50% of the corresponding items had a depth-of-knowledge level below the corresponding objective.

Categorical Concurrence

Two of the four states did not have a sufficient number of assessment items, six or more as used in this analysis, measuring knowledge for more than one-quarter of the standards. The standards and the assessment, in these instances, failed to meet the alignment criterion of categorical concurrence. Standards were not given equal weight on the assessment, as indicated by the number of items related to each standard. Even though most states did not differentiate the emphasis to be placed on one standard over another, the assessments used by the states gave different weightings to standards by varying the number of items measuring context related to the different standards. State D was judged to have the highest proportion of its standards not attaining an acceptable level of categorical concurrence. Of the five different analyses performed for State D, two resulted in half or fewer of the standards attaining the criterion and two resulted in only 62% attaining the criterion. There are some practical reasons why State D had more difficulty meeting categorical concurrence than the other three states. For each content area and grade level, the number of items on the assessment was relatively low compared to the number of standards. State D had the highest number of standards for each grade and content area, while also having among the lowest number of assessment items. State D also had more constructed-response items. This raises the concern of considering each alignment criterion in isolation.

There also may be other reasons why only three or four items are included on an on-demand assessment for specific standards. For example, students' knowledge related to a standard may be more appropriately measured by the teacher in the classroom than on a large-scale assessment. What the analysis of categorical concurrence did uncover was that even with a high number of assessment items being used at a grade level, states have distributed these items unevenly so that one-fourth or more of the standards had less than six items measuring knowledge related to each of these standards.

Depth-of-Knowledge Consistency

The analysis revealed another issue. A high percentage of the state assessments used items at a level of complexity that was below that of the corresponding objectives. This was more of a problem for the alignment in State D than for the other states, in part because State D had a high proportion of its standards at depth-of-knowledge (DOK) levels of strategic thinking (Level 3) and extended thinking (Level 4). Standard statements varied significantly among the states by the amount of content incorporated in one objective and the deepest level of knowledge required to fully meet the standard. For example, State D listed for grade 8 mathematics the following objectives for the Number Sense Standard (V):

V.4 Investigate number forms such as fractions, decimals, and percents, and demonstrate their use in today's society. (DOK: 4)
V.6 Develop, analyze, and explain methods for solving proportions. (DOK: 4)

Reviewers judged both of these objectives to have a depth-of-knowledge level of 4 (Extended Thinking) requiring investigation, development, and applications to society.

In contrast, State B listed as the objective for its grade 8 mathematics Numeration Standard (I):

I.6 Describe the properties of terminating, repeating, and non-repeating decimals and be able to convert fractions to decimals and decimals to fractions. (DOK: 2)

State A stated the following goals and objectives for its grade 6 Numerical and Algebraic Concepts and Operations Standard (I):

I.A Understand and explain how the basic arithmetic operations relate to each other
Apply the distributive property of multiplication over addition and subtraction with integers, fractions, and decimals. (DOK: 2)
I.B Model, explain, and develop reasonable proficiency in operations on whole numbers, fractions, and decimals.
Solve problems that involve addition, subtraction, and/or multiplication with fractions and mixed numbers, with and without regrouping, that include like and unlike denominators of 12 or less and express their answers in the simplest form. (DOK: 3)

Table 2

Summary of Alignment Analysis for Four States in Science and Mathematics

State	Content Area	Grade	Standard	Obj	Depth-of-Knowledge Level of Objectives[a]				Standards with Percent Acceptable Alignment by Criteria[b]				
					1	2	3	4	Item[c]	Cat. Concurr.	Depth	Range	Balance
			#	#	%	%	%	%	#	%	%	%	%
A	Science	3	6	61	16	61	23	0	44	67	83	33	100
		8	6	97	9	56	33	2	70	67	17	33	100
	Mathematics	3	6	94	15	45	26	13	50	67	50	0	100
		6	6	101	10	49	27	14	61	100	100	0	83
B	Mathematics	4	7	61	2	56	34	8	86	100	57	57	86
		8	7	43	0	42	42	16	86	100	71	86	71
		10	4	20	0	35	65	0	70	100	0	100	100
C	Science	4	5	60	8	72	20	0	14	und[d]	und	und	und
		8	5	86	7	77	16	0	14	und	und	und	und
	Mathematics	4	6	107	6	61	31	3	74	100	100	33	83
		8	6	105	14	42	32	12	68	83	83	0	83
D	Science	3	8	86	14	57	20	9	50	38	25	0	100
		7	8	93	11	64	22	3	49	62	50	25	100
		10	8	72	1	56	33	10	46	62	12	12	100
	Mathematics	4	10	56	0	21	41	38	54	90	40	80	90
		8	10	63	0	17	38	44	51	50	40	30	80

[a]1 – Recall
2 – Skill/Concept
3 – Strategic Thinking
4 – Extended Thinking

[b]Categorical Concurrence
Depth-of-Knowledge Consistency
Range-of-Knowledge Correspondence
Balance of Representation

[c]Total number of assessment items

[d]und = undetermined because too few and only sample items were included in the anaysis

What is immediately noticeable are the large differences among the states in the statement of the standards and the articulation of the specific objectives. Even though all of the standards address number, State D places an emphasis on reasoning by labeling the standard as "Number Sense," State B emphasizes more conceptual and procedural knowledge by labeling its standard as "Numeration," and State A extends the conceptual and procedural knowledge to include symbolic manipulation and draws a relationship between number and algebra by labeling the standard as "Numerical and Algebraic Concepts and Operations." The more specific statements of what students are expected to know and to do, referred to here as objectives, represent the differences in the emphases of the standards and vary in the depth-of-knowledge (DOK) levels as rated by the reviewers.

The number of objectives and the range of content addressed by each objective reflect the organization of knowledge and the structure of knowledge incorporated into the standards. As indicated above, the structure-of-knowledge comparability criterion was not used in this analysis, but is necessary to complete a full alignment analysis. The absence of a structure-of-knowledge analysis is one limitation of this study. If one objective had at least one corresponding assessment item, then the objective was considered to be addressed. This was true whether the objective was very broad or narrow. It would be easier for a state to attain an acceptable level on the range-of-knowledge correspondence for a standard with a lower number of objectives. However, incorporating large spans of content in one objective is likely to increase the standard's depth-of-knowledge level, making it more difficult for the standard and assessment to attain an acceptable level on that criterion. One important assumption of this analysis is that the criteria are not independent of each other and multiple criteria need to be incorporated into an analysis to gain a full understanding of alignment. The fact that one state chose to reduce the specificity of its expectations by using fewer objectives will not mean that the alignment will be improved on all criteria.

A review of the general area of geometry provides similar variations among the states:

State D (grade 8):
VI. Geometric and Spatial Sense Standard
2. Explore transformations of geometric figures. (DOK: 4)

State B (grade 8):
II. Geometry Standard
4. Graph on a coordinate plane similar figures, reflections, and translations. (DOK: 2)

State A (grade 6):
IV. Geometry and Spatial Sense Standard
D. Investigate and predict the results of transformations of shapes, figures, and models including slides, flips, and turns. (goal)
1. Identify and describe the results of translations (slides), reflections (flips), rotations (turns), or glide reflections. (DOK: 2)

A review of the standards related to probability and statistics illustrate further the differences among the states and what expectations were incorporated into the standards:

State D (grade 8):
VII. Data Analysis, Probability and Statistics Standard
3. Formulate, predict, and defend positions taken that are based on data collected. (DOK: 4)

State B (grade 8):
VI. Probability and Statistics Standard
1. Collect data involving 2 variables and display on a scatter plot; interpret results (DOK: 3)

State A (grade 6):
VI. Probability and Statistics Standard
E. Make and justify predictions based on collected data or experiments, using technology whenever possible.
1. Evaluate and justify reasoning, inferences, and predictions based on probability and statistics. (DOK: 3)
2. Make inferences and convincing arguments based on probability and statistics and evaluate arguments that are based on probability and statistics. (DOK: 4)

For this illustration of the differences in the statement of expectations among the states, only objectives that addressed nearly the same mathematical topics were included. All three states had other objectives that stated what students should know and do. States varied the most on number and geometry. For geometry and in the area of transformation, State B sought to have students explore the transformation. State B sought to have students graph specific examples of transformations. State A required students to identify and describe specific types of transformations. On probability and statistics, there was less variation among the states on expectations for collecting and analyzing data. Because what states expected their students to know and do varied, different states were challenged more to measure fully whether students attain the desired expectations. Such variation in the level of knowledge required to meet the expectations is reflected in the results of this alignment analysis.

There was no distinct pattern across grade levels and content areas as to which standards and assessments had an acceptable depth-of-knowledge consistency or failed to meet this criterion. State A science had a high degree of depth-of-knowledge consistency for grade 3 but not for grade 8. The reverse was somewhat true for State D science for grades 3 and 7. There was less variation for mathematics than science, but even within mathematics there was variation in the degree that depth-of-knowledge consistency was met for different grades. All of the standards met this criterion for State A grade 6, but only half of the standards met this criterion for State A grade 3.

There was one interesting trend that requires more data for verification. The two times an analysis was done of standards and assessments for grade 10 (State B—mathematics; State D—science), a very low percentage of the standards attained an acceptable criterion of having 50% or more of the items at or above the depth-of-knowledge level of the corresponding objectives. For

BEST COPY AVAILABLE

State B grade 10 mathematics, none of the standards reached this cutoff point. For State D grade 10 science, only 12% of the standards reached this cutoff point. In both of these states, the percentage of standards at grade 10 achieving the criterion was below the percentage of standards that had met this criterion for the middle grade and the elementary grade. This analysis would suggest that expectations are more rigorous for high school, but that the tests did not reflect these expectations and actually had students do less complex tasks.

Some differences were observed between the mathematics and science results that may indicate that reviewers for each of these content areas employed different meanings in their analysis. A higher proportion of mathematics objectives, than science objectives, was assigned to the highest level, Level 4 (Extended Thinking). This may imply that the two groups of reviewers, one in mathematics and one in science, were interpreting the levels differently or it could imply that, in fact, a greater proportion of the mathematics standards than science standards sought to have students extend their thinking.

Of the fourteen complete analyses performed, four of the analyses indicate a high depth-of-knowledge consistency between the standards and the assessments—State A science grade 3 (83%), State A mathematics grade 6 (100%), State C mathematics grades 4 and 8 (100% and 83%, respectively). These four cases illustrate that standards and assessments can be aligned on this criterion.

Range-of-Knowledge Correspondence

States' standards and assessments, along with depth-of-knowledge consistency, achieved the lowest degree of alignment on range-of-knowledge correspondence. The criterion of range-of-knowledge correspondence considered the proportion of the objectives for a standard that had at least one related assessment item. On this criterion, at least 50% of the objectives for a standard had to have a related assessment item or activity to be acceptable. Only State B attained a high degree of range-of-knowledge correspondence for at least two of the analyses completed across all grades and content areas (Table 2). State B grade 8 mathematics had 86% of the standards meet this criterion and grade 10 mathematics had all of its standards meet this criterion. On some of the analyses conducted, this criterion was not met by any of the standards (State A mathematics grades 3 and 6; State C mathematics grade 8; and State D science grade 3).

The attainment of the range-of-knowledge correspondence criterion was a function of the number of objectives and the number of items on the assessment instrument. The states that had more specific statements of expectations and delineated the range of content for a standard by a greater number of objectives required more assessment items to be acceptable. All of the assessments analyzed had an adequate number of items, provided the items were judicially distributed among the objectives so that at least half of the objectives could have at a minimum one item measuring content related to that objective. However, across the objectives of a standard, items were generally clustered among a few of the objectives rather than spanning the full range of objectives. As a consequence, many of the tests were judged to measure students' knowledge of only a small proportion of the full domain of content knowledge specified by the standards.

BEST COPY AVAILABLE

The range-of-knowledge correspondence is very similar to thinking about the domain of all possible assessment activities that could be used to measure students' knowledge of a standard. The goals and objectives that describe in more detail what students are expected to know and do to demonstrate their attainment of the standard serve as more detailed specifications of the standard. Ideally, to measure a student's knowledge of a standard, a test would be created that randomly selected from the domain of all possible items that measure content knowledge related to the standard. A random sample of items would generally incorporate items measuring some knowledge from the full range of the objectives, provided of course there are a sufficient number of items in the sample. In the alignment analysis, all of the objectives for a standard were given equal weight. It was assumed that the test should include at least one item that measured knowledge related to each of at least one-half of the objectives. This is a very lenient requirement. If students are tested on only one-half of a body of knowledge, then there is a large amount of knowledge for which we have no measure of either what a student can or cannot do. The results in this analysis need to be interpreted with an understanding of the underlying assumptions. Since the general knowledge of constructing standards and using standards to drive educational change is relatively new, we simply do not know at this time all of the characteristics a test should assume to measure a set of standards or how a set of standards should be constructed.

There are reasons why the full range of objectives is not assessed. For example, the development of assessment items for some objectives is more difficult than for other objectives. Some states depend on teachers to write the assessment items and then select from the available item pool. Such pools may be weak on items measuring some objectives. Even with clear test specifications, the availability of valid assessment items and activities will limit what objectives can be assessed by a test. Thus, the low alignment on range-of-knowledge correspondence could occur for a number of reasons. The interpretation of this criterion needs to be made in the context of the assessment development process.

Balance of Representation

Most of the assessments and standards analyzed were aligned according to the balance of representation. One explanation for this is that the number of objectives is greater than the number of items and there are few items measuring any one objective. Of fourteen complete analyses conducted, 13 found that 80% or more of the standards and the assessments had an acceptable balance of representation (Table 2). Seven of the analyses indicated that the criterion was fully met by all of the standards. In order for the balance-of-representation criterion to be met, the assessment items had to be nearly evenly distributed among the objectives, for a standard that had at least one corresponding assessment item. The balance-of-representation criterion was met by 71% of the standards for State B grade 8 mathematics, and by 90% of the standards for State D's grade 4 mathematics and 80% for its grade 8 mathematics. One standard common to both States B and D that did not meet the balance-of-representation criterion is related to number and computation—State B Standard VII (Computation) and State D Standard V (Number Sense). For these standards, the assessment instruments tended to over-emphasize one form of computation related to one objective compared to the other objectives for these standards.

Reviewer Agreement in Coding

Reviewers had high agreement in judging whether the alignment between an assessment and standards on a criterion was acceptable or not. The average percentage agreement across all of the standards for a content area and grade level among the reviewers ranged from 71% to 100% (Table 3). Reviewers had the highest agreement on judging whether a standard had 6 or more items, the acceptable level for categorical concurrence. The average overall agreement on the sixteen analyses for categorical concurrence was 94%. Reviewers agreed the least on judging whether the 50% or more of the items related to a standard had a depth-of-knowledge level that was the same as or above that of the corresponding objective within the standard. The average overall agreement on the sixteen analyses for depth-of-knowledge was 83%. Reviewers' agreement on range-of-correspondence (with an acceptable level of 50% or more of the objectives hit) and on balance-of representation (with an acceptable level of the index value of .7 or higher) was, on the average, 91%.

Table 3
Average Percent Across Standards of Reviewers' Agreement on Acceptable Level for Criterion

State	Content	Grade	Reviewer	Criterion			
				Cat. Concurr.	Depth of Knowledge	Range	Balance
			N	Avg %	Avg %	Avg %	Avg %
State A	Science	3	5	87	77	90	97
		8	3	92	75	89	89
	Mathematics	3	4	92	79	92	88
		6	4	100	79	88	88
State B	Mathematics	4	2	100	93	93	93
		8	3	100	81	100	95
		10	2	100	88	100	100
State C	Science	4	5	100	84	100	100
		8	5	96	84	100	100
	Mathematics	4	3	100	100	83	78
		8	4	89	89	100	100
State D	Science	3	6	83	83	88	92
		7	6	88	71	85	92
		10	6	90	79	81	85
	Mathematics	4	7	91	84	87	72
		8	7	93	77	74	84
Average				94	83	91	91

Summary of Findings on Alignment Criteria

This analysis identified specific ways that a state's standards and assessments were aligned. State standards and assessments varied in meeting the cutoff points among the four criteria. This supports the contention that the relationship between standards and assessments can vary on different dimensions and that a number of criteria are needed to judge their alignment. Depth-of-knowledge consistency and range-of-knowledge correspondence were the two criteria that were achieved by the lowest proportion of standards and assessments. In these cases, too many of the assessment items were at a level of knowledge below that of the objective the item was to measure and too few of the objectives for a standard had related items on the assessment.

States varied in the degree to which their assessment and standards were aligned. State B had high alignment for three grades in mathematics, except on depth-of-knowledge consistency. State C had high alignment for two grades in mathematics except on range-of-knowledge correspondence. The alignment for States A and D varied by content area (science and mathematics) and by grade level. A lower percentage of standards and assessments for science met the categorical concurrence criterion than for mathematics. This indicates that the science assessments did not have an adequate number of items measuring each standard for a larger proportion of the standards than for mathematics. Most of the standards and assessments analyzed did have an acceptable balance-of-representation.

Findings: The Process for Studying Alignment

An important goal for this project was to develop valid procedures for performing alignment analyses of standards and assessments. This was the first time that an analysis was done using these alignment criteria. As important as providing states with some feedback on the alignment between their standards and assessments was the refinement of procedures for studying alignment. The steps in the process employed in this alignment analysis included:

1. Identify criteria and acceptable levels
2. Identify expectations and assessments for the state
3. Develop the coding matrix for each content area and grade level
4. Train reviewers
5. Have reviewers code assessment items in relation to objectives
6. Enter data codes onto spreadsheet
7. Analyze data
8. Prepare summary data tables
9. Report results

Reviewers at the four-day institute were not given extensive training at the beginning of the institute. This was done for a reason. The panel of experts was to actively engage in defining and clarifying the process. This took place as reviewers analyzed the standards and assessments state by state and grade by grade. At the end of each analysis they discussed their findings and the process. At the end of the four days, they engaged in a debriefing that clarified very specific ways the process could be improved. A number of recommendations were made for improving the process and for studying the alignment between standards and assessments.

Reviewers and Their Training

Reviewers need to include those who are content-area experts. Reviewers reported they had to consider deeply the knowledge required for a student to successfully answer an item. This concentrated analysis was required even if the assessment item was at a level of recall (Level 1) or skill/concept (Level 2). In addition to content-area experts, it was also helpful to include on the panel those knowledgeable about a state's standards and assessments. At different times during an analysis, reviewers had questions about the context regarding how the standards and assessments were to be used and how these documents were developed. For example, determining the depth-of-knowledge level of some assessment items depended on knowing what materials and equipment are normally available to students, such as calculators and measuring devices. It would be almost impossible for an external panel to synthesize all of the necessary information about the standards and assessments without a significant amount of time and effort. Having persons with this knowledge participating in the analysis was very helpful and made it possible to answer reviewers' questions on the spot.

Reviewers need some training and calibration before doing any of the coding. It was not necessary to spend an excessive amount of time in training, particularly if there were three or more reviewers. Averaging the coding results among the reviewers helped account for differences in coding. At the summer institute, reviewers improved in their agreement as they continued to code. Because only marginal statistics were used in the analysis (totals for an objective and standard), it was not necessary for reviewers to have exact item-by-item agreement. Adequate training could consist of having reviewers code the first three or four items together, followed by a discussion of the results. This process could then be repeated for the next three or four items until reviewers reach consistency in their coding and understand the procedures. Training also could be done by having the reviewers code pre-selected anchor assessment items. Checks of reviewer agreement can be incorporated at different points until there is sufficient evidence that they are interpreting the depth-of-knowledge levels, the standards, and the assessment items in the same way.

Reviewers began each analysis by assigning a depth-of-knowledge level to each objective for a standard. This served two purposes: First, it provided a means for comparing the depth-of-knowledge levels for objectives with the assessment items. Its second purpose was to better acquaint the reviewers with the objectives, goals, and standards. Because reviewers had to reach consensus on the depth-of-knowledge level for each objective, this forced them to consider each objective in some detail.

A critical aspect of the training is for the reviewers to understand the four depth-of-knowledge levels. This requires the reviewers to discuss the verbs and other signals they can use to make the distinctions among the levels. Reviewers did become tainted or less effective over time as a result of rating more than one state. Reviewers had difficulty retaining in their minds how objectives were worded when they coded documents from two states one after the other. This caused reviewers to become more fatigued and to make more mistakes. Recalibration was important. Some consideration should be given to completing the analysis of the standards and assessments from a given state in one day. Reviewers did experience some interference in their thinking in trying to recall and locate objectives that matched assessment items. It was difficult

for reviewers to retain in their minds the content objectives they had to consider after coding over a long time.

Coding process. Reviewers needed help and direction in identifying the central piece of knowledge measured by, and the main purpose for, an assessment task, item, or activity. Because reviewers were able to code an assessment item as related to multiple objectives, reviewers initially differed greatly in how they interpreted what was being assessed by an item. Any item that included a number could be coded as related to number sense. If the standards included "process" standards, then some reviewers developed their own rules that included each item that had to be coded as related to one process standard and one content standard. For example, some of the states had a standard on communication. Initially, some reviewers coded nearly all of the items as related to a communication standard because it required students to read the question. In order to gain stronger agreement among themselves, reviewers developed decision rules, such as coding on the central content knowledge being measured by an item and limiting the number of multiple hits to two or three.

Another decision rule had to be developed on how to consider the context or situation for an assessment item or activity. Some assessment items in both mathematics and science are embedded in a specific context or situation. How reviewers interpreted the context had an impact on what objective they assigned an item to and the depth-of-knowledge level they assigned to the item. For example, some objectives required students to apply their knowledge to a "real-world" problem. Initially, some reviewers coded any story problem as a real-world problem. After being confronted with the need to make some distinctions in the interpretation of context, three ways were identified for addressing contextually embedded assessment items.

1. Superfluous context. The context was only superficial and did not significantly affect students' demonstration of their content knowledge, which was the main intent of the item.

2. Integral context. The context was integral to the assessment item and students' understanding of the content would be different outside of the included context.

3. Partial context. The context is separate, but essential for students to successfully demonstrate the knowledge being measured. For example, students may have to read data from a table or a menu in order to find a value for a routine computation problem. Whereas the computation problem is context free, the student had to abstract information from a simulated situation.

Levels for Determining Depth of Knowledge

Interpreting and assigning depth-of-knowledge levels to both objectives within standards and assessment items is an essential requirement of this alignment analysis. The panel of reviewers agreed that four levels were an adequate number for the purpose of comparing the standards with the assessments. However, Level 2 (Skill/Concept) was used most frequently and did not distinguish among a large number of items and standards. As applied in this analysis, Level 2 was interpreted very broadly—from performing simple procedures to implementing somewhat complex procedures. In contrast, Level 1 (Recall) was frequently interpreted very narrowly—strict recall of memorized facts and information—rather than including other routinized work,

such as using a memorized algorithm. The analysis helped the reviewers to clarify how they used the different levels:

Level 1 (Recall) includes the recall of information such as a fact, definition, term, or a simple procedure, as well as performing a simple algorithm or applying a formula. That is, in mathematics a one-step, well-defined, and straight algorithmic procedure should be included at this lowest level. In science, a simple experimental procedure including one or two steps should be coded as Level 1. Other key words that signify a Level 1 include "identify," "recall," "recognize," "use," and "measure." Verbs such as "describe" and "explain" could be classified at different levels depending on what is to be described and explained.

Level 2 (Skill/Concept) includes the engagement of some mental processing beyond an habitual response. A Level 2 assessment item requires students to make some decisions as to how to approach the problem or activity, whereas Level 1 requires students to demonstrate a rote response, perform a well-known algorithm, follow a set procedure (like a recipe), or perform a clearly defined series of steps. Key words that generally distinguish a Level 2 item include "classify," "organize," "estimate," "make observations," "collect and display data," and "compare data." These actions imply more than one step. For example, to compare data requires first identifying characteristics of the objects or phenomenon and then grouping or ordering the objects. Some action verbs, such as "explain," "describe," or "interpret" could be classified at different levels depending on the object of the action. For example, if an item required students to explain how light affects mass by indicating there is a relationship between light and heat, this was considered a Level 2. Interpreting information from a simple graph, requiring reading information from the graph, also is a Level 2. Interpreting information from a complex graph that requires some decisions on what features of the graph need to be considered and how information from the graph can be aggregated is a Level 3. Caution is warranted in interpreting Level 2 as only skills because some reviewers will interpret skills very narrowly, as primarily numerical skills, and such interpretation excludes from this level other skills such as visualization skills and probability skills, which may be more complex simply because they are less common. Other Level 2 activities include explaining the purpose and use of experimental procedures; carrying out experimental procedures; making observations and collecting data; classifying, organizing, and comparing data; and organizing and displaying data in tables, graphs, and charts.

Level 3 (Strategic Thinking) requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. In most instances, requiring students to explain their thinking is a Level 3. Activities that require students to make conjectures are also at this level. The cognitive demands at Level 3 are complex and abstract. The complexity does not result from the fact that there are multiple answers, a possibility for both Levels 1 and 2, but because the task requires more demanding reasoning. An activity, however, that has more than one possible answer and requires students to justify the response they give would most likely be a Level 3. Other Level 3 activities include drawing conclusions from observations; citing evidence and developing a logical argument for concepts; explaining phenomena in terms of concepts; and using concepts to solve problems.

Level 4 (Extended Thinking) requires complex reasoning, planning, developing, and thinking most likely over an extended period of time. The extended time period is not a distinguishing

factor if the required work is only repetitive and does not require applying significant conceptual understanding and higher-order thinking. For example, if a student has to take the water temperature from a river each day for a month and then construct a graph, this would be classified as a Level 2. However, if the student is to conduct a river study that requires taking into consideration a number of variables, this would be a Level 4. At Level 4, the cognitive demands of the task should be high and the work should be very complex. Students should be required to make several connections—relate ideas *within* the content area or *among* content areas—and have to select one approach among many alternatives on how the situation should be solved, in order to be at this highest level. Level 4 activities include designing and conducting experiments; making connections between a finding and related concepts and phenomena; combining and synthesizing ideas into new concepts; and critiquing experimental designs.

Reviewers faced other issues in deciding on the depth-of-knowledge level to assign both objectives and assessment items. Certain verbs needed clarification, such as "research," "investigate," and "demonstrate." These verbs could be interpreted in different ways, making it more difficult for the reviewers to designate a specific depth-of-knowledge level. For example, sometimes the word "research" was intended to mean that students were expected to look for a term in an encyclopedia and, at other times, "research" indicated that students were to conduct a very complex study. It was difficult to assign a level of complexity to a standard by just using key words without further clarification of the underlying intent of the word. Sometimes this required going beyond the statement of the objective, seeking an example, or considering how the word was used at other grade levels or elsewhere in the documents. In a few cases, it was not possible to be entirely sure how a word was being used. Having input from someone from the state who was knowledgeable about the full intent of the standards was helpful in these situations. There also was an interaction between content and depth-of-knowledge that would influence what depth-of-knowledge level was assigned to an assessment item. An item that requires the recall of a complex or abstract concept is at a higher depth-of-knowledge level than one that only requires the student to recall a simple fact. In addition, the format of an assessment item was a confounding factor in assigning the depth-of-knowledge level to an item. Both a multiple-choice item and an open-response item could require students to interpret a graph, but the multiple-choice item could be a Level 2 and the open-response item a Level 3. Figure 1 provides an illustration of this. The open-response version of this assessment item has a depth-of-knowledge level of 3 (Strategic Thinking), whereas the fixed-response version has a depth-of-knowledge level of 2 (Skill/Concept). The fixed-response item can be worked by eliminating the given choices.

Coding procedures. Reviewers need some context for interpreting the standards and the assessments. This could include a brief summary of what the intended purposes for the standards and assessments were for the state, a list of supporting documents that express the goals of the curriculum and the assessments, how the assessments are to be scored, how the results are to be reported, and some of the supporting documents. A curriculum framework, in addition to the statement of standards, was helpful in interpreting specific words and the underlying meaning of the standards.

Open-Response Version	Fixed-Response Version
A carpenter makes two types of stools. One type of stool has three legs and one type has four legs. Both stool types use the same style of legs. There are 33 legs. How many of each type of stool can the carpenter make, using all of the legs, to have the greatest number of stools?	A carpenter makes two types of stools. One type of stool has three legs and one type has four legs. Both stool types use the same style of legs. There are 33 legs. How many of each type of stool can the carpenter make, using all of the legs, to have the greatest number of stools? A. 3 with 3 legs and 6 with 4 legs B. 5 with 3 legs and 5 with 4 legs C. 7 with 3 legs and 3 with 4 legs D. 9 with 3 legs and 2 with 4 legs

Figure 1. **Example of two assessment items with the same stem, but rated at different depth-of-knowledge levels (Level 3 for the open-response version, and Level 2 for the fixed response version).**

At most, three categories of expectations were used in this analysis—standards, goals, and objectives. Some states use more than three categories of expectations. For example, in addition to these three, some states also identify performance standards and indicators. It is not always clear what the most specific category of expectations for student learning should be in the analysis. In any analysis, the category of expectations that will constitute the level of analysis needs to be clearly specified before beginning the analysis. For most states, but not for all, the most specific statement of expectations should be used. However, State C in this analysis reported student attainment of indicators at the middle category of expectations and not at the most specific category. The most specific category of expectations recorded in its standards document included illustrative examples of what students should be able to do, but was not intended to cover the full span of knowledge as expressed in what was called the "indicator."

This analysis assumed that expectations would be expressed in nested categories, with the standard being the most general, then goal, and then objective. It was found in some cases that the sum of the parts did not always represent the whole. For example, all of the objectives for a goal may be coded at one depth-of-knowledge level, say a Level 2. However, the goal incorporating all of those objectives may be coded at a higher level, say a Level 3 or 4. This is one issue that arose, but was not resolved. The problem derives from how standards are written. One benefit from doing additional analyses of standards and assessments, such as an alignment analysis, is to identify these situations. One of the possibilities for addressing such a situation would be for an assessment to include items at a depth-of-knowledge level that are comparable to that of the goal, but not to any of the specified objectives.

A clear procedure is needed for coding assessment activities that do not match any of the objectives or the category of expectations being compared with assessment activities. For

example, sometimes an item measured content knowledge for a goal, but not content knowledge for any of the listed objectives under that goal. In this analysis, when this situation arose, a general objective was inserted and the item was coded as related to this general objective and not to any of the listed objectives. For example, if an item did not match any of the objectives under Living Systems, but it did fit the goal, then the item was coded as related to the general goal heading of Living Systems. This meant that the coding matrix required one row for the more general category—standard or goal—above the level being coded. Some items did not match any of the standards, goals, or objectives. Some procedures are needed to handle the coding of these items. For example, science reviewers judged that an item on birds and shadows did not fit anything because it was not science. A miscellaneous category, in addition to all of the standards, was needed at each grade level to handle these situations.

Some reviewers felt that there should have been a way to code a near match as well as an exact match. Most of the objectives were robust and covered a range of knowledge, not all of which required the same depth-of-knowledge level of understanding. When trying to relate assessment items to such objectives, reviewers found that an item sometimes matched the objective to some degree, but not exactly for the intended grade. Reviewers estimated that they found exact matches for about 10% to 20% of the items, a near match for about 60% to 70% of the items, and no match for the remainder of items. Distinguishing between near matches and exact matches may help the coding process, but it would add to the problems for doing the analyses.

Limitations of the Alignment Analysis

Some limitations to the analysis were revealed both in the coding and in the aggregating of results. Some of these limitations can be addressed in the design of other alignment studies, but some are more inherent in standards and standards-based education. One issue that was revealed was that assessments could include items targeted for more than one grade level. For example, State A's middle grade test included items for both grades 7 and 8. The grade 8 items actually were intended to measure a higher level of student knowledge than stated in the grade 7 standards used in the alignment analysis. This meant there was some misalignment by design. Again, this and other issues can be revealed and incorporated in the reporting of the results if those who are knowledgeable about the full intent and specifications for the standards and assessments are included on the panel of reviewers.

If an analysis involves multiple states in one session, as in this case, the order in which the state analyses are done is important. The differences in the level of specificity, or the gain size, of the standards for the different states is a critical factor. It is better to begin with a state with the most specific standards and objectives and one that has a range of examples than one with standards that are more difficult to decipher. Also, a state with performance indicators will be easier for reviewers to interpret, which will help in their development as effective coders. Of the states included in the review, the order should have been State C (one with clear performance indicators that helped to bridge the gap between a standard and an indicator for assessment), State D (most complex), State A, and State B.

Reviewers were specifically instructed to judge neither the qualities of the assessment items nor the standards, but to focus their attention on judging how the assessment items matched the

expectations for student learning as stated. This was problematic for some of the reviewers. For one state, State D, all of the reviewers found one or more items on the grade 3 science assessment that they felt did not measure science and that they excluded from the analysis. Also, State D science standards expected grade 8 students to have a conceptual and scientific understanding about cells that has only been developed within the last ten years—a conceptualization that some reviewers felt was too advanced for middle school students. Because reviewers were asked to focus on the alignment between expectations and assessments, the results and reports of the quantitative analysis did not represent all of what the reviewers had learned and understood about each state's system. Some method is needed, such as debriefing the reviewers or having reviewers write their impressions of the standards and assessments, for gathering all of the information that the reviewers have gained from the process.

For the lowest level of analysis, in this case the objective, the analysis does not indicate the degree to which items span the entire domain of possible items for that objective. Multiple items coded for an objective could all measure a narrow range of knowledge included in the objective. This could be problematic and needs to be looked at further. For example, whereas one objective required students to describe, create, extend, and form a generalization of a pattern, nearly all of the items reviewed required students only to extend a pattern. Even though the analysis indicated there was alignment, a large number of items coded for that objective only covered a small range of the possible content. The current system distinguishes items only at the objective level. It does not indicate whether the set of items constituted an adequate sampling from the domain of all possible items for that objective. For a more refined analysis, there is a need to include some means for identifying the degree to which assessments cover the full span of knowledge for an objective and are of a high quality.

Setting an Acceptable Level for Alignment Criteria

Even though this analysis used the same acceptable level to judge the alignment between standards and assessments for all four states, this may not be appropriate because the different states had different purposes for the standards and assessments. Some states prepared standards to influence classroom practices. Other states designed their assessments more for accountability purposes. Along with states having different purposes, it is impractical if not impossible to assess all of the important learning goals on an on-demand assessment. Some of the states expected some of the standards and learning objectives to be assessed by teachers in their classrooms. These and other factors need to be taken into consideration in judging how much alignment is "good enough." Because states will vary in how they use standards and assessments, some consideration needs to be given to the acceptable levels by which the criteria should vary by state.

In addition to different purposes, states use different procedures for sampling what items are included on an assessment and for sampling the content knowledge to be assessed on any one test. States recognize that any test is only a sample of all possible items. One strategy states use to address the issue of adequately sampling content is to test different parts of the desired body of knowledge in different years. Over a span of three or four years, different tests will be administered in a series, so that the full range of content knowledge is tested. In this situation, the

entire series of the tests should be included in the alignment analysis rather than the test for only one year.

Recommendations for Improving the Process

The general sequence of steps employed in this analysis (see page 9) was found to work well. As many as seven and as few as two reviewers analyzed the agreement between the standards and assessments for a grade level and content area. The means among the reviewers were reported. Even when only two reviewers analyzed the alignment, they reached strong agreement on most marginal statistics reported. Having at least three reviewers, however, added more assurance to the results. It is recommended that at least three reviewers be used in coding the assessment items. Because of the number of items and the use of multiple hits, it was difficult and too time-consuming to accurately check on reviewer agreement during the coding process. Therefore, using three or more reviewers increased the likelihood that the results would be more stable.

Based on the analysis and comments from the reviewers, the following recommendations were made to improve the alignment analysis:

1. Incorporate more training to enable reviewers to reach a common understanding of the depth-of-knowledge levels. This training does not need to be extensive and could be done with selected items and standards.

2. Add a means for enabling reviewers to comment, or provide their commentary, on the quality of individual standards and assessment items.

3. Provide reviewers further guidelines for identifying what knowledge is measured by an assessment item and what range of knowledge a student is expected to exhibit as expressed in a standard, goal, or objective.

4. Provide reviewers with specific rules and limits for coding an item as being related to more than one objective.

5. Report for each standard the distribution of coded items by the depth-of-knowledge level.

Conclusions

This study verifies that the stated criteria can be effectively used to structure an analysis of the degree that standards and assessments are aligned. The four criteria—categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance of representation—were all successfully applied to analyze the assessments and standards of four states. The criteria were applicable even though the structure of the standards and assessments varied greatly among the states. Based on the analyses performed, clear differences among the states were evident, along with common issues faced by all. A high percentage of standards and assessments across the four states failed to achieve depth-of-knowledge consistency. In general, too high a frequency of items were below the depth-of-knowledge level of the corresponding objectives for there to be alignment. One benefit to doing an analysis based on specific criteria,

BEST COPY AVAILABLE

such as those used in this study, is that specific feedback could be provided to states on what needs to be done to improve alignment. The procedures need to be refined and reviewers need more training as indicated in the previous section. However, the criteria and the underlying structure of the analysis proved to be viable for detecting the degree of alignment between assessments and standards and how alignment can be improved.

BEST COPY AVAILABLE

References

Baker, E. L., Freeman, M., & Clayton, S. (1991). Cognitive assessment of history for large-scale testing. In M.C. Wittrock & E. L. Baker (Eds.), *Testing and Cognition* (pp. 131-153). Englewood Cliffs, NJ: Prentice Hall.

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64,* 723-733.

Cohen, S. A. (1987). Instructional alignment: Searching for a magic bullet. *Educational Researcher, 16* (8), 16-20.

Consortium for Policy Research in Education. (1991). *Putting the pieces together: Systemic school reform* (CPRE Policy Briefs). New Brunswick, NJ: Rutgers, The State University of New Jersey, Eagleton Institute of Politics.

Newmann, F. M. (1993). Beyond common sense in educational restructuring: The issues of content and linkage. *Educational Researcher, 22* (2), 4-13, 22.

Smith, M. S., & O'Day, J. (1991). Systemic school reform. In S.H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing.* Politics of Education Association Yearbook. (1990, pp. 233-267). London: Taylor & Francis.

Spillane, J. P. (1998). State policy and the non-monolithic nature of the local school district: Organizational and professional considerations. *American Educational Research Journal, 35* (1), 35-63.

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. In *Journal of Educational Measurement, 25* (1), 47-55.

U.S. Congress, House of Representatives. (1994). *Improving America's Schools Act.* Conference report to accompany H.R. 6 Report 103-761. Washington, DC: U.S. Government Printing Office.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education.* Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6. Madison, WI: University of Wisconsin.

Webb, N. L. (1999a). *State A: Alignment between standards and assessments in science for grades 3 and 8 and mathematics for grades 3 and 6.* Council of Chief State School Officers and National Institute for Science Education. Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L. (1999b). *State B: Alignment between standards and assessments in mathematics for grades 4, 8, and 10.* Council of Chief State School Officers and National Institute for Science Education. Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L. (1999c). *State C: Alignment between standards and assessments in science for grades 4 and 8 and mathematics for grades 4 and 8.* Council of Chief State School Officers and National Institute for Science Education. Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L. (1999d). *State D: Alignment between standards and assessments in science for grades 3, 7, and 10 and mathematics for grades 4 and 8.* Council of Chief State School Officers and National Institute for Science Education. Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.

Wixcon, K. K., Fisk, M. C., Dutro, E., & McDaniel, J. (1999). *The alignment of state standards and assessments in elementary reading* (draft). A report commissioned by the National Research Council's Committee on Title I Testing and Assessment.

Appendix A

Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education

1 – Content Focus. System components should focus consistently on developing students' knowledge of subject matter. Consistency will be present to the extent components' logic of action and the ends achieved share the following attributes:

A. *Categorical Concurrence.* Agreement in content topics addressed.
B. *Depth-of-Knowledge Consistency.* Agreement in level of cognitive complexity of information required.
C. *Range-of-Knowledge Correspondence.* Agreement in the span of topics.
D. *Structure-of-Knowledge Comparability.* Agreement in what it means to know concepts.
E. *Balance of Representation.* Agreement in emphasis given to different content topics.
F. *Dispositional Consonance.* Agreement in attention to students' attitudes and beliefs.

2 – Articulation Across Grades and Ages. Students' knowledge of subject matter grows over time. All system components must be rooted in a common view of how students develop, and how best to help them learn at different developmental stages. This common view is based on:

A. *Cognitive Soundness Determined by Superior Research and Understanding.* All components build on principles for sound learning programs.
B. *Cumulative Growth in Knowledge During Students' Schooling.* All components are based on a common rationale regarding progress in student learning.

3 – Equity and Fairness. When expectations are that all students can meet high standards, aligned instruction, assessments, and resources must give every student a reasonable opportunity to demonstrate attainment of what is expected. System components that are aligned will serve the full diversity in the education system through demanding equally high learning standards for all students while fairly providing means for students to achieve and demonstrate the expected level of learning. To be equitable and fair, time is required for patterns to form in order to decipher how system components are working in concert with each other. Judging a system on the criterion of equity and fairness will require analysis over a period of time.

4 – Pedagogical Implications. Classroom practice greatly influences what students learn. Other system components, including expectations and assessments, can and should have a strong impact on these practices, and should send clear and consistent messages to teachers about appropriate pedagogy. Critical elements to be considered in judging alignment related to pedagogy include:

A. *Engagement of Students and Effective Classroom Practices.* Agreement among components in a range of learning activities and in what they are to attain.
B. *Use of Technology, Materials, and Tools.* Agreement among components in how and to what ends applications of technology, materials, and tools are to be included.

5 – System Applicability. Although system components should seek to encourage high expectations for student performance, they also need to form the basis for a program that is realistic and manageable in the real world. The policy elements must be in a form that can be used by teachers and administrators in a day-to-day setting. Also, the public must feel that these elements are credible, and that they are aimed at getting students to learn the mathematics and science that are important and useful in society.

Table AS8-4
Summary of Attainment of Acceptable Alignment Level on Four Content Focus Criteria
State A–Grade 8 Science
(Total Number of Assessment Items—60 Multiple Choice, 4 Open-Response, 6 Open-Ended, Total 70 items)

Standards	Alignment Criteria			
	Categorical Concurrence	Depth-of-Knowledge Consistency	Range of Knowledge	Balance of Representation
1. Process Skills	YES	WEAK	NO	YES
2. Plan and Conduct Investigations	NO	UNDETERMINED	NO	YES
3. Area I: Living Things	YES	YES	YES	YES
4. Area II: Earth and Space Systems	YES	NO	YES	YES
5. Area III: Matter and Energy	YES	WEAK	WEAK	YES
6. Area IV: Applications	NO	NO	NO	YES

Single copy price is $7.75. To order copies contact:

CENTER DOCUMENT SERVICE
Wisconsin Center for Education Research
1025 W. Johnson St., Room 242
Madison, WI 53706-1796
608/265-9698

NO PHONE ORDERS. PREPAYMENT REQUIRED FOR ORDERS UNDER $20.00.

Price is subject to change without notice.

National Institute for Science Education
University of Wisconsin–Madison
1025 West Johnson Street
Madison, WI 53706

(608) 263-9250
(608) 262-7428 fax
niseinfo@macc.wisc.edu
http://www.nise.org

Eisenhower National Clearinghouse
43210 000 148209